KLASTERIZĀCIJU RAKSTUROJOŠO PARAMETRU IETEKME UZ DATU ANALĪZES REZULTĀTIEM

IMPACT OF PARAMETERS CHARACTERIZING CLUSTERING ON DATA ANALYSIS RESULTS

Pēteris GRABUSTS

Dr. sc. ing., asoc. prof., Rēzeknes Augstskola
Tālrunis: +371 26593165, e–pasts: peter@ru.lv
Rēzekne, Latvija

Abstract. *Clustering algorithms are used to group some given objects defined by a set of numerical properties in such a way that the objects within a group are more similar than the objects in different groups. All clustering algorithms have common parameters the choice of which characterizes the effectiveness of clustering. The most important parameters characterizing clustering are: metrics (the distance between cluster elements and cluster centre), number of clusters k and cluster validity criteria. The goal of the paper – to perform the evaluation of the validity of metrics' choice, to describe the change with respect to the number of clusters for experimental data purposes and to evaluate the credibility of clustering results. As an input data the table describing the rating of Latvian state higher educational institutions for year 2011 has been used and the goal of the experiment was to show, how by using the clustering methods it is possible to analyze the mentioned data in an alternative way.*
Keywords: *clustering algorithms, metrics, k–means, cluster validity.*

Ievads

Mūsdienās ir uzkrājies liels daudzums datu dažādās zinātnes, uzņēmējdarbības, tautsaimniecības u.c. sfērās un rodas nepieciešamība analizēt tos labākai konkrētās nozares vadīšanai. Bieži uzņēmējdarbības vajadzības stimulē izstrādāt jaunas intelektuālās datu analīzes metodes, kas ir orientētas uz praktisku pielietojumu. Klasteranalīzes kā viena no intelektuālās datu analīzes pamatuzdevumiem mērķis ir neatkarīgu grupu (klasteru) un to raksturlielumu meklēšana analizējamos datos. Šāda uzdevuma atrisināšana ļauj labāk izprast datus, jo klasterizāciju var izmantot praktiski jebkurā pielietojumu jomā, kur nepieciešama datu analīze.

Klasteranalīzes pamatā ir hipotēze par kompaktumu. Tiek pieņemts, ka apmācošās kopas elementi pazīmju telpā atrodas kompaktā veidā. Galvenais uzdevums – formalizēti aprakstīt šos veidojumus. Ir pazīstami daudzi klasterizācijas algoritmi – tādi kā *Isodata*, *FOREL*, k–vidējais (*k–means*) u.c. Visiem klasterizācijas algoritmiem ir kopīgi parametri, kuru izvēle arī raksturo klasterizācijas efektivitāti. Svarīgākie klasterizāciju raksturojošie parametri ir šādi: metrika (klasteru elementu

attālums līdz klastera centram), klasteru skaits k, klasterizācijas ticamības novērtējums.

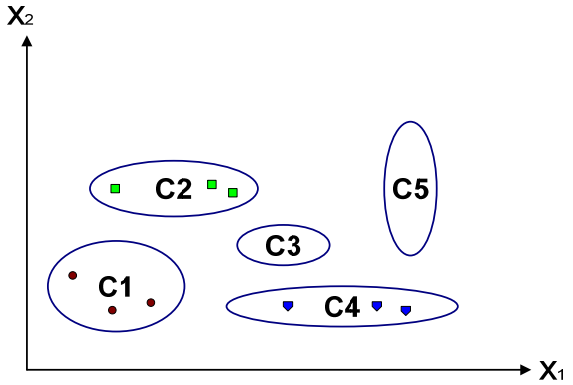
Lai izvērtētu klasterizācijas algoritmu darbības efektivitātes aspektus, autors izvirzīja mērķi – veikt Latvijas valsts dibināto augstskolu reitinga datu par 2011. gadu analīzi. Pētījuma uzdevumi pakārtoti izvirzītajam mērķim:

- veikt metrikas izvēles pamatotības izvērtēšanu;
- raksturot klasteru skaita izmaiņu analizējamajiem datiem;
- novērtēt klasterizācijas rezultātu ticamību.

Pētījuma nolūks bija parādīt, kā ar klasterizācijas metodēm alternatīvā veidā var analizēt šādus datus. Galvenās pētījuma metodes dotajā darbā ir aprakstošā metode, matemātiskā modelēšana un statistiskā analīze. Pētījuma laika periods ir 2011. gads.

1. Klasterizācijas metožu pielietojums datu analīzē

Dažādās pētniecības jomās aktuāls ir jautājums: "Kā organizēt novērojamos datus pārskatāmās struktūrās?". Pastāv viedoklis, ka, atšķirībā no daudzām citām statistiskām procedūrām, vairumā gadījumu klasteranalīzes metodes tiek izmantotas tad, kad nav nekādu hipotēžu attiecībā par klasēm, bet vēl aizvien notiek datu vākšanas etaps. Klasteranalīzes metodes ļauj sadalīt pētāmos objektus "līdzīgu" objektu grupās, ko sauc par klasteriem (6.). Klasterizācijas būtība ir attēlota 1. attēlā, kur divdimensiju telpas objekti sadalīti 5 klasteros.



1. attēls. Divdimensiju objektu telpas sadalījuma klasteros piemērs
Avots: autora izveidots attēls

Klasterizācija atšķiras no klasifikācijas ar to, ka analīzes veikšanai klasterizācijas procesā nav nepieciešamības izdalīt atsevišķu mainīgo grupu. No šī viedokļa klasterizācija tiek traktēta kā „apmācība bez skolotāja" un tiek pielietota pētījumu sākotnējā posmā (9.).

Klasteranalīzi raksturo divas īpatnības, kas atšķir to no citām metodēm:

1) rezultāts atkarīgs no objektu vai to atribūtu dabas, t.i., tie var

būt viennozīmīgi noteikti objekti vai arī objekti ar izplūdušu aprakstu;
2) rezultāts atkarīgs no iespējamās klastera un objektu attiecības klasteros dabas, t.i., jāņem vērā objekta iespējamā piederība vairākiem klasteriem un objekta piederības noteikšana (stingra vai izplūdusī piederība).

Ņemot vērā klasterizācijas svarīgo lomu datu analīzē, objekta piederības jēdziens tika vispārināts uz tādu klašu funkciju, kas nosaka klašu objektu piederību konkrētai klasei.

Izdala divu veidu klases raksturojošās funkcijas:
1) diskrētā funkcija, kas pieņem vienu no divām iespējamajām vērtībām – pieder/nepieder klasei (klasiskā klasterizācija);
2) funkcija, kas pieņem vērtības no intervāla [0,1]. Jo funkcijas vērtības tuvāk 1, jo objekts „vairāk" pieder konkrētai klasei (izplūdusī klasterizācija).

Klasterizācijas algoritmi pārsvarā paredzēti daudzdimensiju datu izlašu apstrādei, kad dati doti tabulas veidā "objekts–īpašība". Tie ļauj grupēt objektus noteiktās grupās, kurās objekti saistīti savā starpā pēc kādas konkrētas kārtulas. Nav svarīgi, kā tiek dēvētas šādas grupas – taksoni, klasteri, klases, galvenais, ka tās pietiekami precīzi atspoguļo šo objektu īpašības. Pēc klasterizācijas datus tālākai analīzei izmanto citas intelektuālās datu analīzes metodes, lai noskaidrotu iegūto likumsakarību būtību un turpmākās izmantošanas iespējas (5.).

Klasterizāciju parasti izmanto datu apstrādes procesā kā pirmo analīzes soli. Tā identificē līdzīgu datu grupas, kas vēlāk var tikt izmantotas datu kopsakarību pētīšanā (3.). Klasteranalīzes process formāli sastāv no šādiem etapiem:

- analīzei nepieciešamo datu savākšana;
- klašu datu (klasteru) raksturojošo lielumu un robežu noteikšana;
- datu grupēšana klasteros;
- klašu hierarhijas noteikšana un rezultātu analīze.

Datu analīzē tradicionāli tiek pielietots k–vidējais klasterizācijas algoritms (2.). Tas minimizē kvalitātes rādītāju, kurš noteikts kā visu punktu, kas pieder klastera apgabalam, attālumu līdz klastera centram (metriku) (1.). Ar metriku šajā kontekstā tiek saprasta distance (attālums) starp klasterā ietilpstošajiem punktiem (7.). Parasti klasterizācijas algoritmos ieejas datu vektors tiek salīdzināts ar citiem vai ar iepriekš noteiktu klastera centru. Attāluma metrika arī nosaka piederību tam vai citam klasterim, tādējādi nosakot likumsakarības daudzdimensiju datu izlasēs – attiecinot ieejas datus tai vai citai klasei jeb klasterim (8.).

Eiklīda attālums (*(Euclidean distance*) ir visbiežāk izmantojamā distance klasterizācijā. Tas ir attālums starp divu punktu koordinātēm daudzdimensiju telpā, kas atbilst tos savienojošā nogriežņa garumam un ko aprēķina pēc formulas (3.):

$$D_{XY} = \sqrt{\sum_{k=1}^{m} (x_{ik} - x_{jk})^2} \quad (1)$$

Manhetenas attālumu (*Manhattan distance*) aprēķina kā koordinātu pāru vērtību starpību absolūto summu (3.):

$$D_{XY} = \sum_{k=1}^{d} | x_{ik} - x_{jk} | \quad (2)$$

Attāluma vērtību var izteikt arī ar Pirsona korelācijas koeficienta (*Pearson correlation coefficient*) palīdzību. Izskaitļotā vērtība atrodas intervālā no –1 līdz 1 un raksturo punktu līdzības pakāpi (3.):

$$r_{ij} = \frac{\sum_{k=1}^{d} (x_{ik} - \bar{x}_i)(x_{jk} - \bar{x}_j)}{\sqrt{\sum_{k=1}^{d} (x_{ik} - \bar{x}_i)^2 \sum_{k=1}^{d} (x_{jk} - \bar{x}_j)^2}}, \text{ kur } \bar{x}_i = \frac{1}{d} \sum_{k=1}^{d} x_{ik} . \quad (3)$$

Kosinusa distance (*Cosine distance*) ir leņķiskā starpība starp diviem datu punktiem (3.):

$$D_{XY} = \cos(\theta) = \frac{X \bullet Y}{\|X\| \|Y\|} = \frac{\sum_{i=1}^{n} X_i \times Y_i}{\sqrt{\sum_{i=1}^{n} (X_i)^2} \times \sqrt{\sum_{i=1}^{n} (Y_i)^2}} \quad (4)$$

Autors savā darbā (4.) pārbaudīja klasiskā klasterizācijas algoritma k–vidējais darbības rezultātus ar dažādām metrikām: Eiklīda attālumu, Manhetenas distanci, Kosinusa distanci un Pirsona korelācijas koeficientu. Eksperimentu gaitā kā k–vidējais klasterizācijas algoritmā klasteru centru noteikšanai secīgi tika izmantotas minētās četras metrikas. Iegūtie rezultāti tika analizēti un tika pārbaudīta klasterizācijas kvalitāte. Tradicionāli klasterizācijas algoritmos izmanto Eiklīda attālumu, taču citas metrikas izvēle atsevišķos gadījumos var būt diskutējama. Tas atkarīgs no risināmā uzdevuma, datu apjoma un sarežģītības. Tika konstatēts, ka klasterizācijas rezultāti visu apskatāmo metriku izmantošanā ir ļoti līdzīgi. Nevienai no izvēlētajām metrikām nebija izšķirīga pārsvara, kas ļautu pasludināt to par labāko.

Algoritms k–vidējais izpildās šādu soļu secībā (9.):

(a) inicializē klasteru centrus w_j (j – nepieciešamo klasteru skaits uzdevuma risināšanai);
(b) grupē visus apmācības izlases punktus ap tuvākā klastera centru,

t.i., katru punktu x_i saista ar klasteru j^*, kuram

$$\| x_i - w_{j^*} \| = \min_j \| x_i - w_j \| ;$$

(c) izskaitļo jaunos klasteru centrus, t.i., visiem w_j izskaitļo :

$$w_j = \frac{1}{m_j} \sum_{x_i \in \text{klasterim } j} x_i \quad , \text{ kur } m_j - \text{klasterim j piederošo punktu skaits;}$$

(d) atkārto (b) soli tik ilgi, kamēr iterāciju laikā nemainās klasteru centru vērtības.

Algoritma darbības rezultātā tiek noteikti galīgie klasteru centri w_j, ievērojot nosacījumu, ka attālumu kvadrātu summai starp visiem punktiem, kas pieder grupai j, un klastera centru ir jābūt minimālai.

Būtisks jautājums k–vidējais algoritma realizēšanā ir klasteru skaita un sākotnējo centru noteikšana. Vienkāršākajos uzdevumos pieņem, ka *apriori* ir zināms klasteru skaits un par sākotnējām m klasteru centru vērtībām tiek piedāvāts ņemt apmācošās kopas pirmos m punktus.

Par algoritma k–vidējais priekšrocībām var uzskatīt popularitāti, lielo efektivitāti un procedūras vienkāršību. Bet gadījumā, kad objektu izvietojums ir neviendabīgs, algoritms var nedot labus rezultātus. Tad ir jāmaina parametri (klasteru skaits) un atkal jāmēģina atkārtot algoritma darbības. Par trūkumu tiek uzskatīts tas, ka algoritms nav universāls.

2. Klasterizācijas rezultātu ticamības novērtējums

Darba izstrādes laikā aktualizējās jautājums par klasterizācijas kvalitātes kritērijiem, t.i., skaitliska kritērija noteikšana, lai varētu novērtēt klasterizācijas rezultātu.

Literatūrā apskatīti trīs kritēriji klasterizācijas ticamības novērtēšanā: ārējais kritērijs, iekšējais kritērijs un relatīvais kritērijs (9.). Raksta ierobežotā apjoma dēļ turpmāk tiks apskatīts tikai klasterizācijas ārējais ticamības indekss.

Ja dota datu izlase X un klasteru struktūra C tiek iegūta no X klasterizācijas algoritma darbības rezultātā, tad ārējais kritērijs salīdzina iegūto klasteru struktūru C ar specificēta sadalījuma P kategorijām, kas atspoguļo potenciālo klasteru struktūru, ko varētu iegūt no X. Datu punktu pārim x_i un x_j no izlases X pastāv četri stāvokļi, kā x_i un x_j izvietojas klasteru struktūrās C un P (9.):

- Stāvoklis 1: x_i un x_j pieder līdzīgiem C klasteriem un līdzīgām P kategorijām.
- Stāvoklis 2: x_i un x_j pieder līdzīgiem C klasteriem, bet atšķirīgām P kategorijām.
- Stāvoklis 3: x_i un x_j pieder atšķirīgiem C klasteriem, bet līdzīgām P kategorijām.

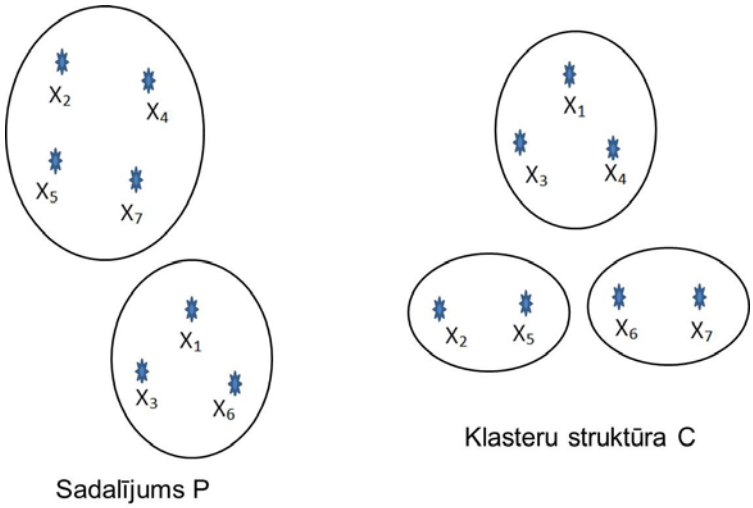
- Stāvoklis 4: x_i un x_j pieder atšķirīgiem C klasteriem un atšķirīgām P kategorijām.

Punktu pāri visiem četriem stāvokļiem tiek apzīmēti ar a, b, c un d. Tā kā punktu pāru kopskaits ir n(n–1)/2, tad var iegūt izteiksmi:

$$M = a + b + c + d = \frac{n(n-1)}{2}, \quad (5)$$

kur n ir datu izlases punktu skaits.

Ārējā indeksa izskaitļošanas demonstrēšanas nolūkā pieņem, ka dota datu izlase ar 7 punktiem. 2. attēlā parādīts punktu izvietojums specificētā sadalījumā P un klasterizācijas rezultātā iegūtajā klasteru struktūrā C.



2. attēls. Punktu pāri četru stāvokļu demonstrēšanai (9.)

Attēlā redzams, ka, piemēram, stāvoklim 1 atbilst tikai divi gadījumi – (x_1 un x_3) un (x_2 un x_5). Šie punkti vienlaikus pieder līdzīgiem C klasteriem un līdzīgām kategorijām sadalījumā P, t.i., atrodas vienā un tajā pašā klasterī sadalījumā P un vienā un tajā pašā klasterī struktūrā C. Visiem četriem stāvokļiem atbilstošie punktu pāri parādīti 1. tabulā.

1.tabula

Datu punktu gadījumu stāvokļi (9.)

Stāvoklis	Datu punktu pāri	Kopā
a	x_1 un x_3; x_2 un x_5	2
b	x_1 un x_4; x_3 un x_4; x_6 un x_7	3
c	x_1 un x_6; x_2 un x_4; x_2 un x_7; x_3 un x_6; x_4 un x_5; x_4 un x_7; x_5 un x_7	7
d	x_1 un x_2; x_1 un x_5; x_1 un x_7; x_2 un x_3; x_2 un x_6; x_3 un x_5; x_3 un x_7; x_4 un x_6; x_5 un x_6	9

Kad C un P ir noteikti, var izvēlēties kādu no daudzajiem klasterizācijas kvalitātes kritērijiem (9.). Dotajā pētījumā klasterizācijas

kvalitātes kritēriji tika novērtēti ar Randa indeksa un Huberta indeksa palīdzību.

Randa indeksu skaitļo pēc šādas formulas (3.):

$$R = \frac{a+d}{M} \tag{6}$$

Huberta indekss (3.):

$$\Gamma = \frac{1}{M}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} X_{ij}Y_{ij} \tag{7}$$

Abu indeksu vērtības ir robežās no 0 līdz 1. Lielāka indeksa vērtība liecina par lielāku līdzību starp C un P.

3. Latvijas augstskolu 2011. gada reitinga analīze ar klasterizācijas palīdzību

3.1. Reitinga dati

Lai varētu novērtēt klasterizāciju raksturojošo parametru ietekmi uz klasterizācijas rezultātiem, tika veikts pētījums, kurā tika izmantoti Latvijas valsts dibināto augstskolu reitinga tabula par 2011. gadu (10.).

Augstskolu starptautiskie reitingi kļūst arvien populārāki. Pastāv dažādas metodoloģijas augstskolu reitingu noteikšanā.

Reitings *Webometrics Ranking* ranžē vairāk nekā 20 000 augstākās mācību iestādes pasaulē (11.). Reitings balstās tikai uz Internetā pieejamās informācijas par mācību iestādi. Tiek izmantoti četri galvenie indikatori: 10% no ranga vērtības sastāda augstskolas atpazīstamība Google meklētājā; 50% – ārējo saišu skaits uz augstskolas mājas lapu; 10% – akadēmiskās un publicēšanās aktivitātes dažādu datņu formātā Google meklētājā (.doc, .pdf, .ppt); 30% – elektronisko publikāciju skaits no Google Scholar (2007–2011) un dati no Scimago SIR (2003–2010).

Šajā reitingā LU ierindota 822. vietā, RTU – 915. vietā, LLU – 3119. vietā, TSI – 3436. vietā, RA – 3659. vietā.

Reitings *SCImago Institutions Rankings* ranžē 3042 augstākās mācību iestādes pasaulē un balstās uz datiem par augstskolas zinātniskajām aktivitātēm (12.). Četri indikatori ietver informāciju par publikāciju skaitu (galvenokārt SCOPUS), zinātniskās sadarbības rādītājus, augsta līmeņa publikāciju skaitu utt. No Latvijas augstskolām šeit minēta LU (1565. vietā) un RTU (2794. vietā).

Reitings *QS World University Rankings* veido 500 pasaules vadošo augstskolu izlasi (14.). Tiek izmantoti 6 indikatori: 40% – akadēmiskā reputācija; 10% – darba devēju reputācija; 20% – zinātnisko darbu citējamība; 20% – studentu īpatsvars; 5% – ārzemju studentu īpatsvars; 5% – starptautisko fakultāšu īpatsvars. Latvijas augstskolas šajā reitingu tabulā nav pārstāvētas.

Reitings *The Times Higher World University Ranking (THE)* veido 400 pasaules vadošo augstskolu izlasi (13.). Tiek izmantoti 13 indikatori, kas grupēti 5 grupās: 30% – apmācības vide; 30% – pētnieciskais darbs; 30% – citējamība; 2,5% – inovācijas; 7,5% – ārzemju sakari. Latvijas augstskolas šajā reitingu tabulā nav pārstāvētas.

Latvijas augstskolu reitinga izveidē pamatā ņemta THE metodoloģija un vērtēšanas kritēriji jeb indikatori ir šādi:

- I1– studējošo un akadēmiskā personāla skaita attiecība;
- I2– absolventu īpatsvars;
- I3– akadēmiskā personāla pamatdarbā ar Dr. grādu īpatsvars (starp visām augstskolām);
- I4– akadēmiskā personāla pamatdarbā ar Dr. grādu īpatsvars (konkrētajā augstskolā)
- I5– akadēmiskā personāla pamatdarbā īpatsvars;
- I6– akadēmiskā personāla vecuma struktūra (30 – 50 g. veco īpatsvars);
- I7– ārzemju studentu īpatsvars;
- I8– publikāciju skaits uz akadēmiskā personāla 1 vienību;
- I9– izglītības kvalitāte (izcila un laba);
- I10– augstskolas popularitāte/ atpazīstamība.

Augstskolu reitinga rezultējošie dati parādīti 2. tabulā.

2.tabula

Latvijas valsts dibināto augstskolu 2011. gada reitinga dati (10.)

Augst–skola	I1	I2	I3	I4	I5	I6	I7	I8	I9	I10	Vieta
LU	63	13,5	150	63	44,5	46	39,5	200	184	100	1
RTU	45	11	124,5	60	50	36	21,5	166	200	100	2
RSU	53	12,5	57	57	50	38	44	100	198	99	3
DU	45	10,5	33	59	49,5	55	2,5	138	82	96	4
LLU	35	11	49,5	60	26	39	0	38	148	99	5
REA	9	13,5	3	85	11	77	19	20	88	94	6–7
LNAA	0	50	7,5	69	50	67	0	0	82	89	6–7
VeA	18	10	7,5	38	42,5	38	0,5	58	96	87	8
LMāA	6	13,5	3	10	50	52	0	0	154	97	9
RPIVA	68	14,5	12	46	38	47	0,5	0	68	88	10–12
LMūA	5	12	4,5	12	49	42	0,5	0	162	94	10–12
BA	50	18	4,5	33	26,5	31	0,5	0	120	94	10–12
LSPA	23	11	10,5	51	43,5	32	0	0	106	95	13
RA	63	12	10,5	36	39,5	65	2	0	48	86	14
ViA	30	11	4,5	27	34,5	66	0	20	80	83	15–17
LKuA	8	10,5	4,5	25	43	47	1	0	122	93	15–17
LJA	21	4,5	3	33	30,5	13	0,5	0	148	97	15–17
LiepU	39	13	13,5	49	21	38	0	0	72	93	18

Turpmākajos pētījumos tika izmantotas šo indikatoru skaitliskās vērtības. Netika ņemti vērā ģeogrāfiskie, sociālie un politiskie aspekti, kā arī iegūtā vieta reitingu tabulā.

3.2. Klasterizācijas rezultāti

Pētījumā tika izdarīts mēģinājums sagrupēt augstskolas ar klasterizācijas algoritma k–vidējais palīdzību un pārliecināties, vai šāds sadalījums atbilst matemātiski izskaitļotajai augstskolas vietai reitinga tabulā. Eksperimentālā daļa tika veikta MatLab vidē un iegūtie klasteri tika salīdzināti ar SPSS klasterizācijas rezultātiem. Secīgi izvēloties klasteru skaitu robežās no 2 līdz 8 un pielietojot klasterizācijas algoritmu k–vidējais, tika iegūti attiecīgie klasteri un to komponenti (3. tab.).

3.tabula

Klasterizācijas rezultātā iegūtie klasteri un to komponenti

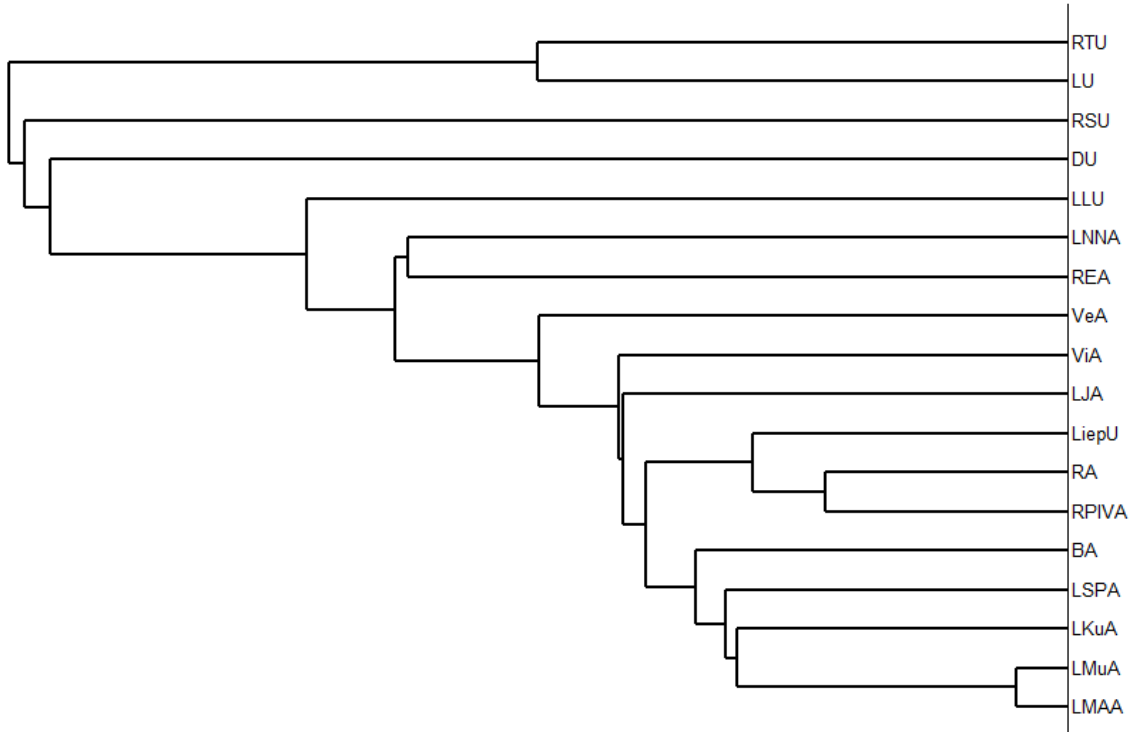
Klasteru skaits	Augstskolas klasteros							
2	LU RTU RSU	Pārējās						
3	LU RTU	RSU	Pārējās					
4	LU RTU	RSU	DU	Pārējās				
5	LU RTU	RSU	DU	RPIVA RA ViA LiepU	Pārējās			
6	LU RTU	RSU	DU	RPIVA RA ViA LiepU	REA LNAA	Pārējās		
7	LU RTU	RSU	DU	RPIVA RA ViA LiepU	REA LNAA	LLU VeA BA LSPA	LMāA LKuA LMūA LJA	
8	LU RTU	RSU	DU	RPIVA RA ViA LiepU	REA	LLU VeA BA LSPA	LMāA LKuA LMūA LJA	LNAA

Avots: autora aprēķinu rezultāts

Tabulā redzams, ka pirmajos trijos klasteros ietilpstošās augstskolas reitinga tabulā atrodas reitinga tabulas augšējā daļā. Tāpat secināms, ka 5,6,7 un 8 klasteru gadījumā algoritma darbības rezultātā izskaitļoto četru klasteru sastāvs ir nemainīgs. Atšķirības sākas ar piekto klasteri. Arī ar SPSS iegūtie rezultāti ir līdzīgi.

Klasteru vizualizācijai bieži izmanto hierarhisko klasterizāciju (2.). Ja divi klasteri nokļūst vienā grupā līmenī k un tie paliek kopā arī augstākos līmeņos, tad šādu grupēšanu sauc par hierarhisko klasterizāciju. Jebkurai hierarhiskai grupēšanai pastāv atbilstoša kokveida struktūra, ko sauc par dendrogrammu, kas parāda kā grupējas

klasteri. Hierarhiskās klasterizācijas rezultātā iegūtā augstskolu reitinga tabulas datu dendrogramma parādīta 3. attēlā.



3. attēls. Augstskolu reitinga datu dendrogramma
Avots: autora aprēķinu rezultāts

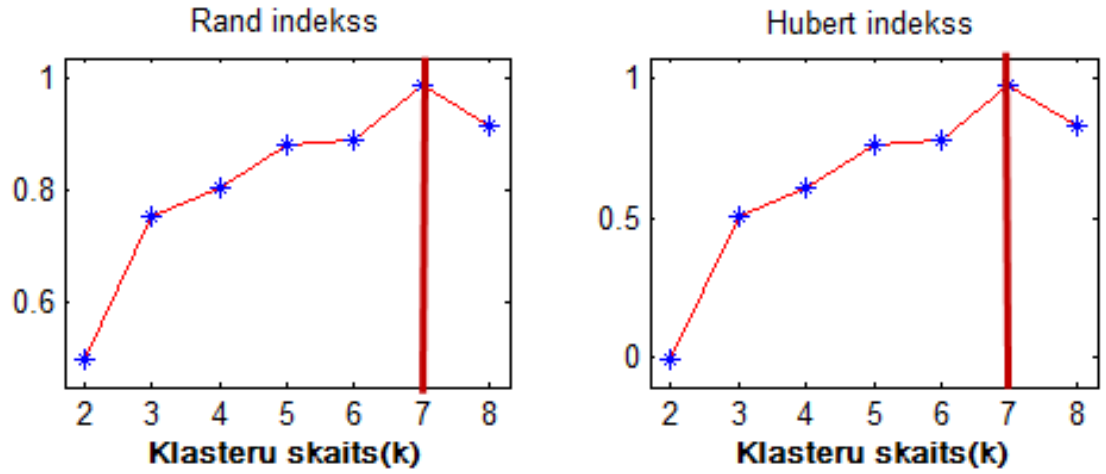
Dendrogrammas analīze liecina, ka rezultējošie klasteri būtiski neatšķiras no k–vidējais klasterizācijas algoritma darbības rezultātā iegūtajiem klasteriem (3. tab.).

3.3. Klasterizācijas ticamības analīze

Lai pārbaudītu veiktās klasterizācijas ticamību, tika izskaitļoti kvalitātes rādītāji – Randa un Huberta indeksi astoņiem klasteriem. Klasteru struktūra C (secīgi ar klasteru skaitu robežās no 2 līdz 8 klasteriem) tika salīdzināta ar specificētajiem sadalījumiem P, kas satur dažādus potenciālos klasterus. Piemēram, ja klasteru struktūra C sastāv no diviem klasteriem, tad ņem sadalījumu P ar trim klasteriem, izskaitļo a, b, c, d vērtības un izskaitļo kvalitātes indeksu struktūrai C un identificēšanas kļūdu. Tad ņem sadalījumu P ar četriem klasteriem utt. Izskaitļotās kopējās klasterizācijas kļūdas bija šādas:

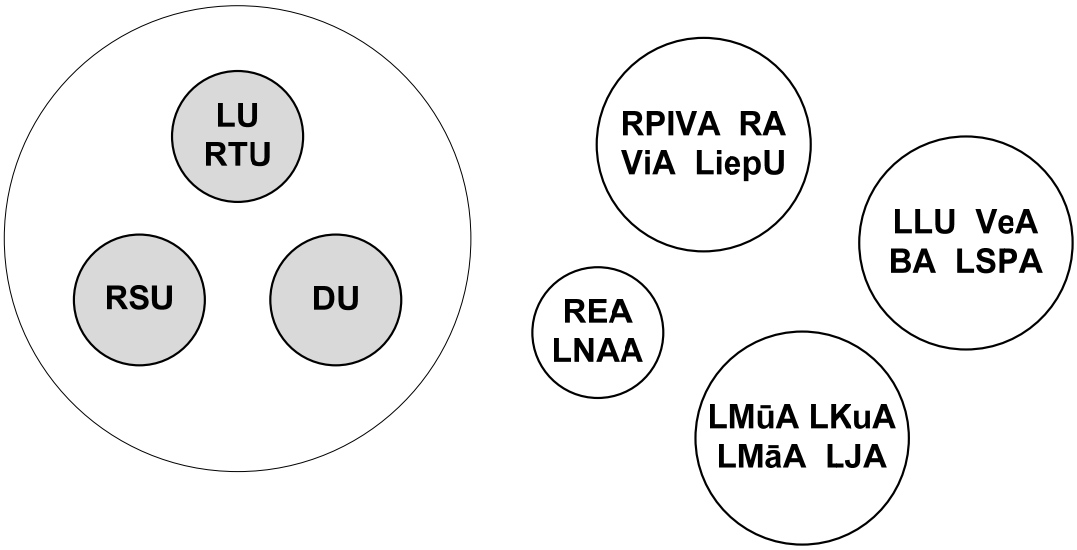
- 2 klasteriem – 5,56 %;
- 3 klasteriem – 50%;
- 4 klasteriem – 55,6%;
- 5 klasteriem – 55,6%;
- 6 klasteriem – 38,89%;
- 7 klasteriem – 33,33%;
- 8 klasteriem – 50%.

Starp visām C struktūrām vismazākā kļūda ir 7 klasteru gadījumā, t.i., 7 klasteru struktūra šajā gadījumā ir visoptimālākā. 4. attēlā parādīti izskaitļotie Randa un Huberta indeksi.



4. attēls. Randa un Huberta indeksi septiņu klasteru gadījumā
Avots: autora aprēķinu rezultāts

Tādējādi ir konstatēts, ka doto datu izlasi vislabāk raksturo 7 klasteru struktūra (3. tab.). Ņemot vērā publiskajā telpā izskanējušās runas par augstskolu restrukturizācijas nepieciešamību, no matemātiskā viedokļa raugoties 7 izskaitļotos optimālos klasterus varētu turpināt apvienot, iegūstot, piemēram, „superklasteri" ar LU, RTU, RSU un DU. Šāds hipotētisks sadalījums parādīts 5. attēlā.



5. attēls. Augstskolu sadalījums septiņu klasteru gadījumā
Avots: autora izveidots attēls

Var secināt, ka klasterizācijas kvalitāti raksturojošie indeksi ir ļoti lietderīgi klasterizācijas algoritmu darbības rezultātu analīzē. Ar to

palīdzību var izvēlēties optimālu klasteru struktūru gadījumos, kad datu sadalījums klasteros sākotnēji nav noteikts.

Secinājumi un priekšlikumi

Klasteranalīzes kā viena no intelektuālās datu analīzes pamatuzdevumiem mērķis ir neatkarīgu grupu (klasteru) un to raksturlielumu meklēšana analizējamos datos. Šāda uzdevuma atrisināšana ļauj labāk izprast datus, jo klasterizāciju var izmantot praktiski jebkurā pielietojumu jomā, kur nepieciešama eksperimentālo vai statistisko datu analīze.

Visiem klasterizācijas algoritmiem ir kopīgi parametri, kuru izvēle arī raksturo klasterizācijas efektivitāti. Svarīgākie klasterizāciju raksturojošie parametri ir šādi: metrika (klasteru elementu attālums līdz klastera centram), klasteru skaits k un klasterizācijas ticamības novērtējums.

Lai izvērtētu klasterizācijas algoritmu darbības efektivitātes aspektus, tika izvirzīts mērķis– veikt Latvijas valsts dibināto augstskolu reitinga datu par 2011. gadu analīzi. Pētījuma uzdevumi bija pakārtoti izvirzītajam mērķim:

- veikt metrikas izvēles pamatotības izvērtēšanu;
- raksturot klasteru skaita izmaiņu analizējamajiem datiem;
- novērtēt klasterizācijas rezultātu ticamību.

Pētījuma nolūks bija parādīt, kā ar klasterizācijas metodēm alternatīvā veidā var analizēt šādus datus. Pētījuma laikā tika konstatēts, ka metrikas izvēle būtiski neiespaido klasterizācijas kvalitāti. Daudz lielāka nozīme bija klasteru skaita izvēlei. Reitinga datu klasterizācijā un pēc tās veiktās klasterizācijas indeksa izskaitļošanas par optimālāko tika izvēlēta klasteru struktūra ar septiņiem klasteriem. Pētījuma eksperimentālās daļas rezultāti liecina, ka augstskolas šajos klasteros iedalītas pēc to „tuvības" mēra, ko nosaka indikatoru vērtības. Tāpat tika secināts, ka vietu reitinga tabulā būtiski ietekmē indikatora I8 (publikāciju skaits) vērtība.

Tāda veida datu analīze ar klasterizācijas palīdzību var tikt uzskatīta kā papildus līdzeklis tradicionālajām datu apstrādes procedūrām, bet tās rezultāti ir rūpīgi jāanalizē.

Analizējot atsauksmes presē par augstskolu reitingu korektumu, tika secināts, ka galvenie iebildumi pret 2011. gada augstskolu reitingu ir šādi:

- nav korekti salīdzināt studējošo ārvalstnieku skaitu privātās un valsts augstskolās;
- nav korekti apgalvot: „jo lielāka augstskola, jo tā ir kvalitatīvāka";

- šī gada topu nevar salīdzināt ar iepriekšējo, jo abu gadu topos ir izmantoti atšķirīgi vērtēšanas kritēriji;
- šogad izmantota zinātnisko publikāciju datu bāze „SCOPUS". Tā pilnībā neatspoguļojot zinātniskās publikācijas, piemēram, humanitārajās zinātnēs;
- reitingā vairāk punktu ieguvušas tās augstskolas, kurās ir liels studentu skaits uz vienu mācībspēku;
- atsevišķi esot jāvērtē vecākas un jaunākas augstskolas, galvaspilsētā esošās un reģionālās augstākās izglītības iestādes.

Tā kā esošā reitinga aprēķina metodoloģija tiek pamatīgi kritizēta no augstskolu pārstāvju puses, būtu lietderīgi tomēr izstrādāt metodoloģiju, kas apmierinātu lielāko augstskolu daļu.

Izmantotā literatūra un avoti

1. AGRAWAL, R. et al. *Efficient similarity search in sequence databases*. Proc. 4th Int. Conf. On Foundations of Data Organizations and Algorithms, Chicago.1993. pp. 69–84.
2. EVERITT, B. *Cluster analysis*. Edward Arnold, London, 1993.
3. GAN, G. et al. *Data clustering: Theory, algorithms and applications*. ASA–SIAM series on Statistics and Applied Probability, SIAM, Philadelphia, ASA, Alexandria, VA, 2007.
4. GRABUSTS, P. *Distance Metrics Selection Validity in Cluster Analysis*. RTU zinātniskie raksti. 5. sēr., Datorzinātne. 49. sēj. 2011. 72.–77. lpp.
5. HAN, J. et al. *Geographic Data Mining and Knowledge Discovery*. Taylor and Francis, 2001. 372 pages.
6. KAUFMAN, L., ROUSSEEUW, P. *Finding groups in data. An introduction to cluster analysis*. John Wiley & Sons, 2005.
7. LI, M. et al. *The similarity metric*. IEEE Transactions on Information Theory, vol.50, No. 12, 2004. pp.3250–3264.
8. VITANYI, P. *Universal similarity*. ITW2005, Rotorua, New Zealand, 2005.
9. XU, R., WUNVH, D. *Clustering*. John Wiley & Sons, 2009. pp. 263–278.
10. KUZMINA, I. *Augstskolu vērtēšana uzkurina kaislības* [tiešsaiste]. Laikraksta "Latvijas Avīze" publikācija [atsauce 2012.g. 15.feb.]. Pieejas veids: http://la.lv/index.php?option=com_content&view=article&id=314680:augstskolu–vrtana–uzkurina–kaislbas&catid=124:aktuli&Itemid=146
11. *Rank of Universities of Latvia* [tiešsaiste]. Ranking Web of World Universities [atsauce 2012.g. 15.feb.]. Pieejas veids: http://www.webometrics.info/rank_by_country.asp?country=lv
12. *SIR World Report 2011*[tiešsaiste]. SCImago Institutions Rankings [atsauce 2012.g. 15.feb.]. Pieejas veids: http://www.scimagoir.com/
13. *Top 400 World Universities* [tiešsaiste]. The Times Higher World University Ranking [atsauce 2012.g. 15.feb.]. Pieejas veids: http://www.timeshighereducation.co.uk/world–university–rankings/2011–2012/top–400.html

14. *QS World University Rankings 2011/2012* [tiešsaiste]. QS Top Universities [atsauce 2012.g. 15.feb.]. Pieejas veids: http://www.topuniversities.com/university–rankings/world–university–rankings/2011

Summary

The goal of cluster analysis as one of the main tasks of intellectual data analysis is the search of independent groups (clusters) and their characteristics in the data under consideration (analysis). Solving such task allows to better understand the data, since clustering can be used in practically all sorts of applications where analysis of experimental or statistical data is required.

Traditionally k–means clustering algorithm has been used in the research. The following could be mentioned as its advantages: popularity, high efficiency and simplicity of the procedure.

All clustering algorithms have common parameters, the choice of which characterizes the efficiency of clustering. Some of the most important parameters describing the clustering are:

- metric (the distance between cluster elements and cluster centre);
- number of clusters k;
- evaluation of clustering validity.

The goal of the research was to perform validity evaluation with respect to the choice of metrics, to describe changes in the number of clusters with respect to the experimental data and to evaluate the credibility of clustering results. Rating table of the Latvian state higher educational institutions for year 2011 has been used as input data and the goal of the experiment was to show how by applying clustering methods the mentioned data can be analyzed in an alternative way.

During the research an attempt has been made to group the higher educational institutions with the help of k–means clustering algorithm and to verify whether such division corresponds to the rate of certain higher educational institution in the rating table calculated mathematically.

Traditionally, Euclidean distance is used in clustering algorithms, however, the choice of another metrics in certain cases may be justified. It depends on the task under consideration, the amount of data and level of complexity. It has been discovered that the results of clustering by using all metrics under consideration are very similar. Neither of the chosen metrics (Euclidean distance, Manhattan distance, Pearson correlation coefficient, Cosine distance) dominated so that it could be named as the best choice.

Much more important was selection of the number of clusters. In the clustering of rating data and after the performance of the calculation of clustering index, cluster structure with seven clusters has been chosen as the most optimal structure. In the given research the criteria for clustering efficiency have been evaluated with the help of Rand index and Hubert index. It has been concluded that indices characterizing the quality of clustering are very useful for analyzing the results of clustering algorithms' performance. With their help it is possible to choose an optimal cluster structure in cases when data division in clusters has not been initially defined.

Such type of data analysis with the help of clustering can be considered as additional means to traditional data processing procedures and its' results shall be carefully analyzed.